Method of Least Squares

Least Squares Regression

Linear Regression

- Fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

  $y = a_0 + a_1 x + e$

  a_1- slope

  a_0- intercept

  e- error, or residual, between the model and the observations

Criteria for a "Best" Fit/

- Minimize the sum of the residual errors for all available data:

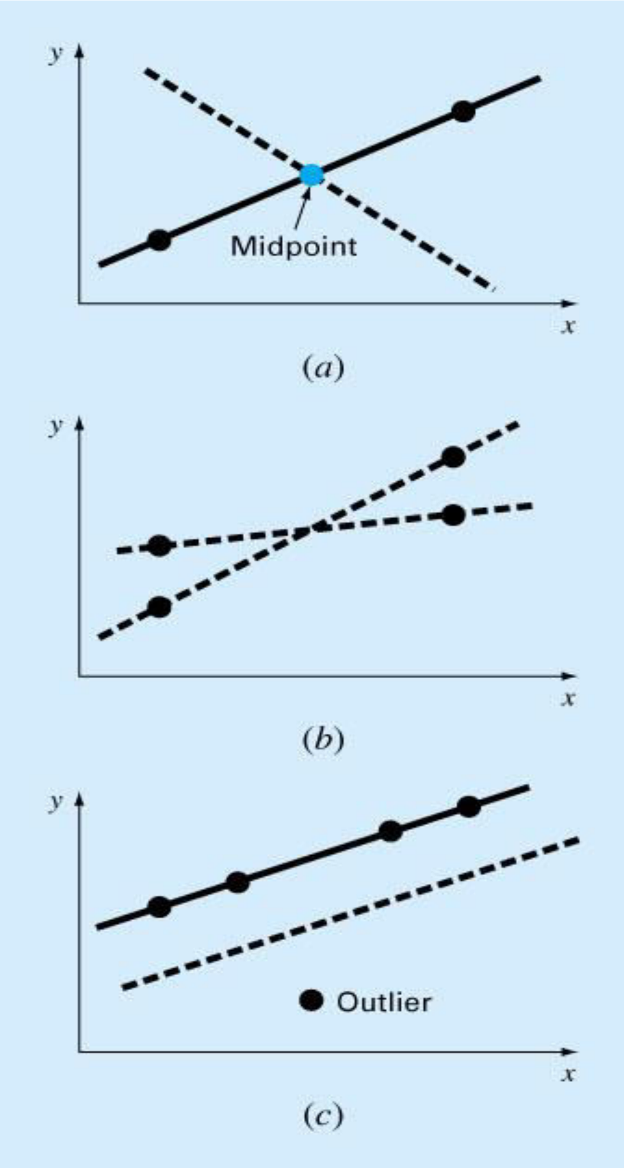
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_o - a_1 x_i)$$

n = total number of points

- However, this is an inadequate criterion, so is the sum of the absolute values

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$

Figure



- Best strategy is to minimize the sum of the squares of the residuals between the measured y and the y calculated with the linear model:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i,\text{measured}} - y_{i,\text{model}})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- Yields a unique line for a given set of data.

Least-Squares Fit of a Straight Line/

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum [(y_i - a_o - a_1 x_i) x_i] = 0$$

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

$$0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$$

$$\left.\begin{aligned} &\sum a_0 = n a_0 \\ &n a_0 + \left(\sum x_i\right) a_1 = \sum y_i \end{aligned}\right\}$$

Normal equations, can be solved simultaneously

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Mean values

$$a_0 = \bar{y} - a_1 \bar{x}$$

Figure :

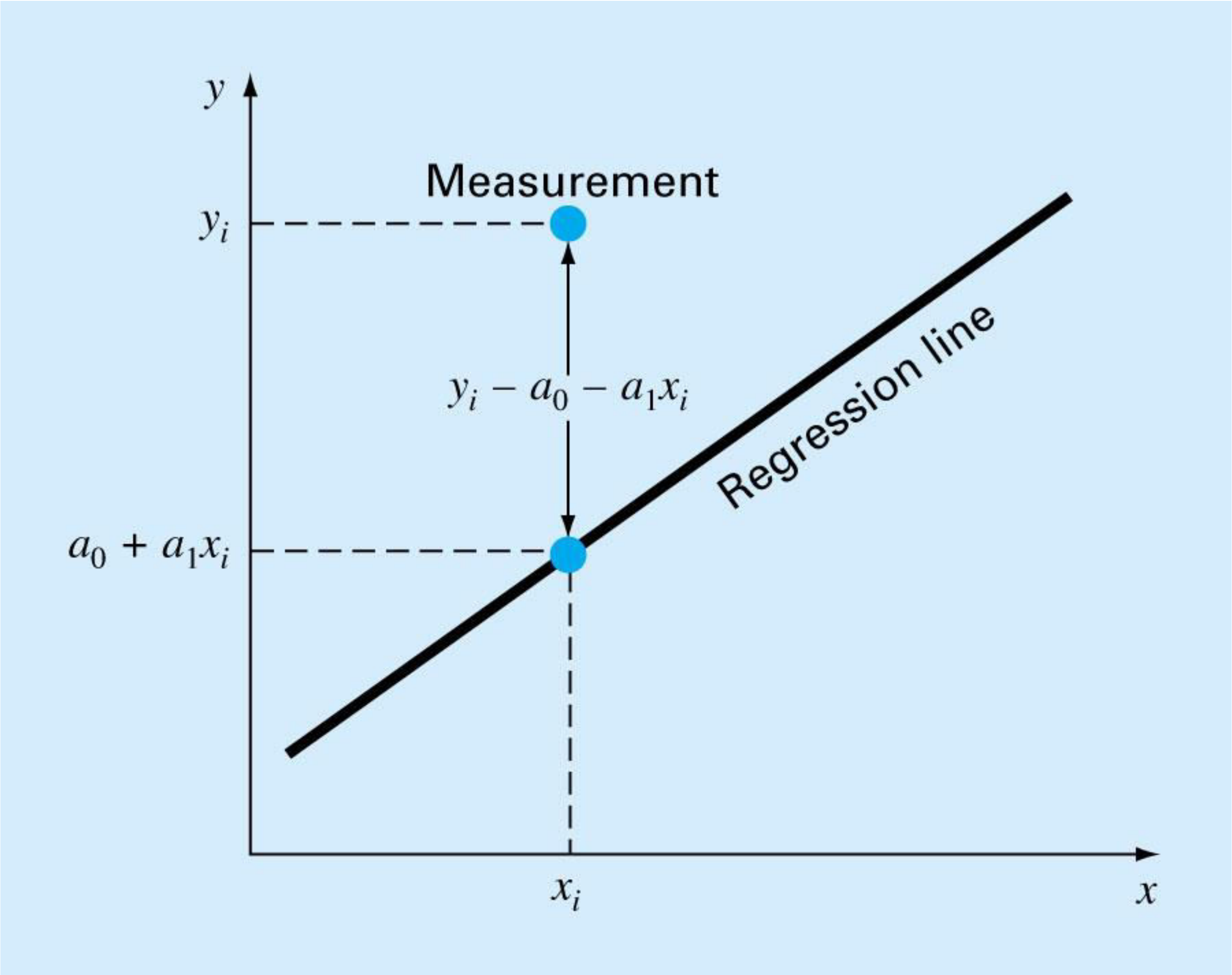


Figure :

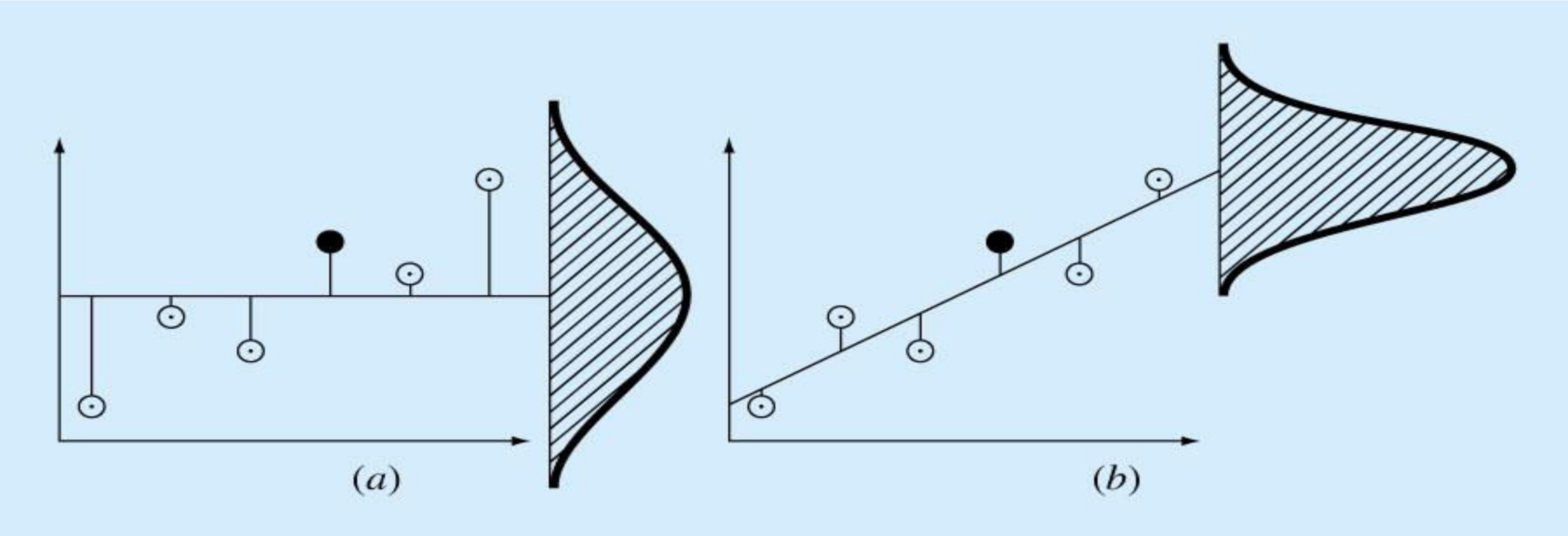
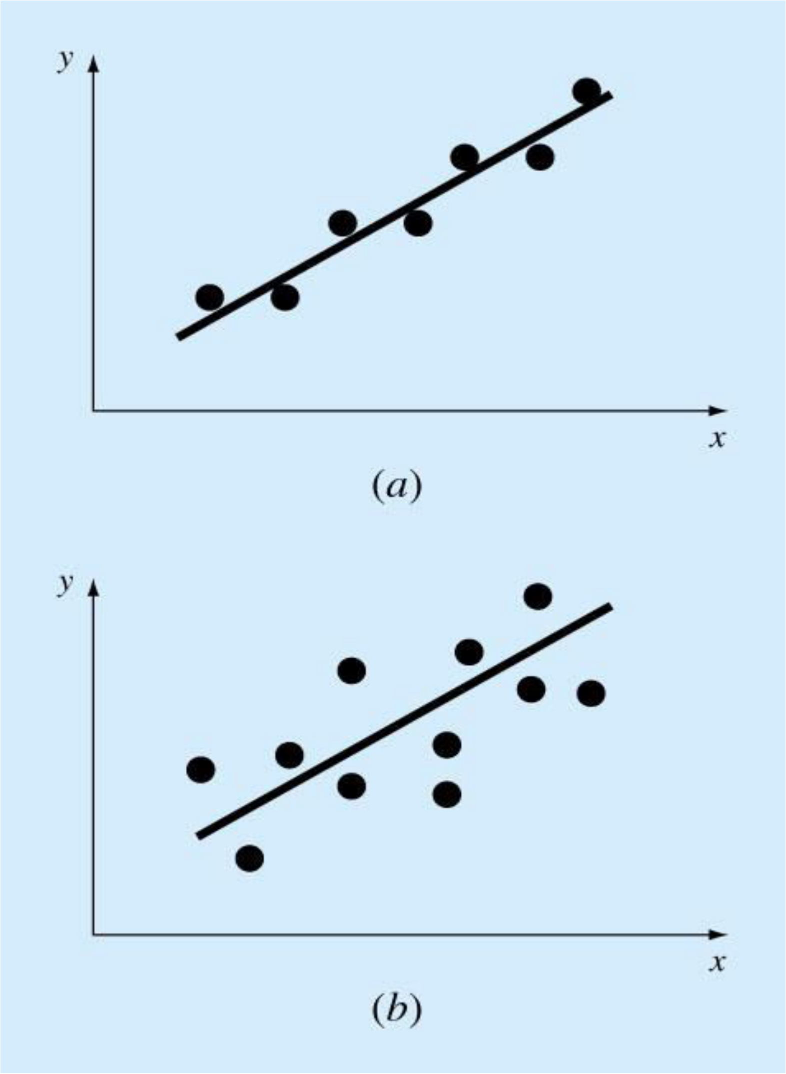


Figure:



“Goodness” of our fit/

If

- Total sum of the squares around the mean for the dependent variable, y, is S_t
- Sum of the squares of residuals around the regression line is S_r
- S_t-S_r quantifies the improvement or error reduction due to describing data in terms of a straight line rather than as an average value.

$$r^2 = \frac{S_t - S_r}{S_t}$$

r^2-coefficient of determination

Sqrt(r^2) – correlation coefficient

- For a perfect fit

  S_r=0 and $r=r^2=1$, signifying that the line explains 100 percent of the variability of the data.
- For $r=r^2=0$, $S_r=S_t$, the fit represents no improvement.

Polynomial Regression

- Some engineering data is poorly represented by a straight line. For these cases a curve is better suited to fit the data. The least squares method can readily be extended to fit the data to higher order polynomials .

General Linear Least Squares

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$z_0,\ z_1,\ \ldots,\ z_m$ are $m+1$ basis functions

$$\{Y\} = [Z]\{A\} + \{E\}$$

$[Z]$ – matrix of the calculated values of the basis functions at the measured values of the independent variable

$\{Y\}$ – observed valued of the dependent variable

$\{A\}$ – unknown coefficients

$\{E\}$ – residuals

$$S_r = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} a_j z_{ji} \right)^2$$

Minimized by taking its partial derivative w.r.t. each of the coefficients and setting the resulting equation equal to zero